QASR: QCRI Aljazeera Speech Resource
A Large Scale Annotated Arabic Speech Corpus

Hamdy Mubarak,[1] Amir Hussein,[2] Shammur Absar Chowdhury,[1] and Ahmed Ali[1]
[1] Qatar Computing Research Institute, HBKU, Doha, Qatar
[2] Kanari AI, California, USA
info@arabicspeech.org, https://arabicspeech.org/

Abstract

We introduce the largest transcribed Arabic speech corpus, QASR[1], collected from the broadcast domain. This multi-dialect speech dataset contains $2,000$ hours of speech sampled at 16kHz crawled from Aljazeera news channel. The dataset is released with lightly supervised transcriptions, aligned with the audio segments. Unlike previous datasets, QASR contains linguistically motivated segmentation, punctuation, speaker information among others. QASR is suitable for training and evaluating speech recognition systems, acoustics- and/or linguistics- based Arabic dialect identification, punctuation restoration, speaker identification, speaker linking, and potentially other NLP modules for spoken data. In addition to QASR transcription, we release a dataset of 130M words to aid in designing and training a better language model. We show that end-to-end automatic speech recognition trained on QASR reports a competitive word error rate compared to the previous MGB-2 corpus. We report baseline results for downstream natural language processing tasks such as named entity recognition using speech transcript. We also report the first baseline for Arabic punctuation restoration. We make the corpus available for the research community.

1 Introduction

Research on Automatic Speech Recognition (ASR) has attracted a lot of attention in recent years (Chiu et al., 2018; Watanabe et al., 2018). Such success has brought remarkable improvements in reaching human-level performance (Xiong et al., 2016; Saon et al., 2017; Hussein et al., 2021). This has been achieved by the development of large spoken corpora: supervised (Panayotov et al., 2015; Ardila et al., 2019); semi-supervised (Bell et al., 2015; Ali et al., 2016); and more recently unsupervised (Valk and Alumäe, 2020; Wang et al., 2021) transcription. This work enables to either reduce Word Error Rate (WER) considerably or extract metadata from speech: dialect-identification (Shon et al., 2020); speaker-identification (Shon et al., 2019); and code-switching (Chowdhury et al., 2020b, 2021).

Natural Language Processing (NLP), on the other hand values large amount of textual information for designing experiments. NLP research for Arabic has achieved a milestone in the last few years in morphological disambiguation, Named Entity Recognition (NER) and diacritization (Pasha et al., 2014; Abdelali et al., 2016; Mubarak et al., 2019). The NLP stack for Modern Standard Arabic (MSA) has reached very high performance in many tasks. With the rise of Dialectal Arabic (DA) content online, more resources and models have been built to study DA textual dialect identification (Abdul-Mageed et al., 2020; Samih et al., 2017).

Our objective is to release the first Arabic speech and NLP corpus to study spoken MSA and DA. This is to enable empirical evaluation of learning more than the word sequence from the speech. In our view, existing speech and NLP corpora are missing the link between the two different modalities. Speech poses unique challenges such as disfluency (Pravin and Palanivelan, 2021), overlap speech (Tripathi et al., 2020; Chowdhury et al., 2019), hesitation (Wottawa et al., 2020; Chowdhury et al., 2017), and code-switching (Du et al., 2021; Chowdhury et al., 2021). These challenges are often overlooked when it comes to NLP tasks, since they are not present in typical text data.

In this paper, we create and release[2] the largest corpus for transcribed Arabic speech. It comprises of $2,000$ hours of speech data with lightly supervised transcriptions. Our contributions are: (*i*)

[1] QASR قصر in Arabic means “Palace”. The acronym stands for: QCRI Aljazeera Speech Resource.

[2] Data can be obtained from: https://arabicspeech.org/qasr

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 2274–2285
August 1–6, 2021. ©2021 Association for Computational Linguistics

aligning the transcription with the corresponding audio segments including punctuation for building ASR systems; (*ii*) providing semi-supervised speaker identification and speaker linking per audio segments; (*iii*) releasing baseline results for acoustic and linguistic Arabic dialect identification and punctuation restoration; (*iv*) adding a new layer of annotation in the publicly available MGB-2 testset, for evaluating NER for speech transcription; (*v*) sharing code-switching data between Arabic and foreign languages for speech and text; and finally, (*vi*) releasing more than 130M words for Language Model (LM).

We believe that providing the research community with access to multi-dialectal speech data along with the corresponding NLP features will foster open research in several areas, such as the analysis of speech and NLP processing jointly. Here, we build models and share the baseline results for all of the aforementioned tasks.

1.1 Related work

The CallHome task within the NIST benchmark evaluations framework (Pallett, 2003), released one of the first transcribed Arabic dialect dataset. Over years, NIST evaluations provided with more dialectal - mainly in Egyptian and Levantine dialects, as part of language recognition evaluation campaign. Projects such as GALE and TRANSTAC (Olive et al., 2011) program, released more than 251 hours of Arabic data, including the first spoken Iraqi dialect among others. These datasets exposed the research community to the challenges of spoken dialectal Arabic and motivated to design competition to handle dialect identification, dialectal ASR among others (see Ali et al. (2021) for details).

The following datasets are released from the Multi-Genre Broadcast MGB challenge: *(i)* MGB-2 (Ali et al., 2016) – this dataset is the first milestone towards designing the first large scale continuous speech recognition for Arabic language. The corpus contains a total of 1,200 hours of speech with lightly supervised transcriptions and is collected from Aljazeera Arabic news channel span over many years. *(ii)* MGB-3 (Ali et al., 2017) – focused on only Egyptian Arabic broadcast data comprises of 16 hours. *(iii)* MGB-5 (Ali et al., 2019) – consists of 13 hours of Moroccan Arabic speech data. In addition, the CommonVoice[3] Arabic dataset, from the CommonVoice project, provides 49 hours of modern standard Arabic (MSA) speech data.[4]

Table 1: Comparison between MGB-2 *vs* QASR.

	MGB-2	QASR
Hours	1,200	2,000
Dialects	MSA, GLF, LEV, NOR, EGY	
Segmentation	Influenced by silence and segment length	Linguistically and acoustically motivated
Transcription	Lightly supervised	Lightly supervised
Punctuation	–	✓
Code-Switching	–	✓
Possible Turn-Ending	–	✓
Speaker Names	✓	✓(+ normalised names)
Speaker Gender	–	✓(2000 speakers) covers ≈82% data
Speaker Country	–	✓Manually annotated in testset
NER	–	✓Manually annotated in testset

Unlike MGB-2, QASR dataset is the largest multi-dialectal corpus with linguistically motivated segmentation. The dataset includes multi-layer information that aids both speech and NLP research community. QASR is the first speech corpora to provide resources for benchmarking NER, punctuation restoration systems. For close comparison between MGB-2 *vs* QASR, see Table 1.

2 Corpus Creation

2.1 Data Collection

We obtained Aljazeera Arabic news channel's archive (henceforth AJ), spanning over 11 years from 2004 until 2015. It contains more than 4,000 episodes from 19 different programs. These programs cover different domains like politics, society, economy, sports, science, etc. For each episode, we have the following: (*i*) audio sampled at 16KHz; (*ii*) manual transcription, the textual transcriptions contained no timing information. The quality of the transcription varied significantly; the most challenging were conversational programs in which overlapping speech and dialectal usage was more frequent; and finally (*iii*) some metadata.

For better evaluation of the QASR corpus, we reused the publicly available MGB-2 (Ali et al., 2016) testset as it has been manually revised, coming from the same channel, thus making this testset ideal to evaluate the QASR corpus. It is worth noting that we ensure that the MGB-2 dev/test sets

[3] `https://commonvoice.mozilla.org/en/datasets`

[4] Reported on June 2021.

Item	Description
Hours	10
Episodes	17. Average episode duration = 34 min
Segments	8, 014
Words	69, 644. Unique words = 15, 754
Speakers	111
Males	87 (78%)
Females	13 (11%)
Variety	MSA: 78%:, Dialectal Arabic: 22%
Countries	Top 5 countries are: (based on dialectal segments) EG: 18%, SY: 11%, PS: 11%, DZ: 8%, SD: 7%
Genre	Top 5 topics are: Politics: 69%, Society: 9%, Economy: 8%, Culture/Art: 4%, Health: 3%

Table 2: *Description of the updated MGB-2 testset*

Item	Count	Notes
Hours	2, 041	
Episodes	3, 545	Average episode duration = 32 min.
Segments	1.6M	. Average segment duration = 4 sec 84% of segments are [2-6] sec . Average segment len = 9 words 80% of segments have [5-11] words
Words	14.3M	Unique words = 360K
Speakers	27, 977	Unique speakers = 11, 092
Males	1, 171	1.2M segments (69%)
Females	68	99K segments (6%)

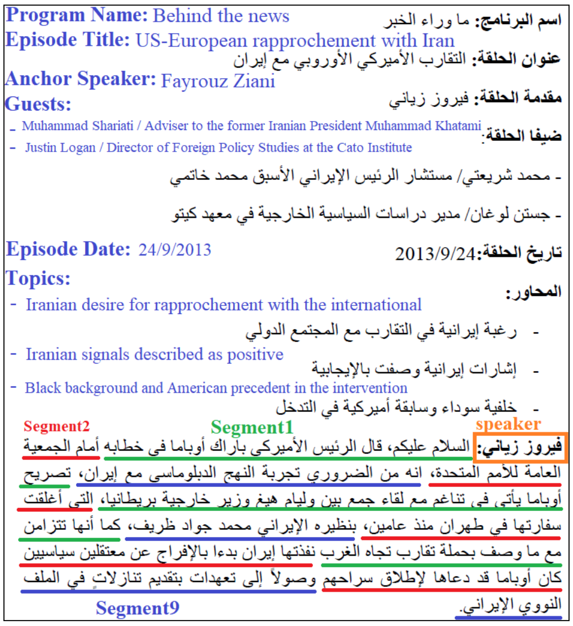
Table 3: *QASR Corpus Statistics*

are not included in QASR corpus, so they can be used to report progress on the Arabic ASR challenge. We have also enriched the MGB-2 testset with manually annotated speaker information like country[5], gender of the speakers, along with NER information and used it to evaluate our baselines.

Moreover, we apply topic classification and dialect identification. Our models achieved an overall accuracy of 96% and 88% respectively, which have been measured on internal testsets also created from Aljazeera news articles. More details can be found in ASAD demo paper (Hassan et al., 2021). Table 2 gives a rough estimate about distributions in the updated MGB-2 testset.

2.2 Metadata Information

Most of the recorded programs have the following metadata: program name, episode title and date, speaker names and topics of the episode. Majority of metadata information appear in the beginning of the file. However, some of them are embedded inside the episode transcription. Figure 1 shows a sample input file from Aljazeera. One of the main challenges is the inconsistency in speaker names, e.g. *Barack Obama* appeared in 9 different forms (*Barack Obama*, *Barack Obama/the US President*, *Barack Obama/President of USA*, *Barck Obama* (typo), etc.). The list of guest speakers and episode topics are not comprehensive, with many spelling mistakes in the majority of metadata field names and attributes. To overcome these challenges, we applied several iterations of automatic parsing and extraction followed by manual verification and standardization.



Program Name: Behind the news
اسم البرنامج: ما وراء الخبر
Episode Title: US-European rapprochement with Iran
عنوان الحلقة: التقارب الأميركي الأوروبي مع إيران
Anchor Speaker: Fayrouz Ziani
مقدمة الحلقة: فيروز زياني
Guests:
- Muhammad Shariati / Adviser to the former Iranian President Muhammad Khatami
- Justin Logan / Director of Foreign Policy Studies at the Cato Institute

ضيفا الحلقة:
- محمد شريعتي/ مستشار الرئيس الإيراني الأسبق محمد خاتمي
- جستن لوغان/ مدير دراسات السياسية الخارجية في معهد كيتو

Episode Date: 24/9/2013
تاريخ الحلقة:2013/9/24
Topics:
المحاور:
- Iranian desire for rapprochement with the international
- رغبة إيرانية في التقارب مع المجتمع الدولي
- Iranian signals described as positive
- إشارات إيرانية وصفت بالإيجابية
- Black background and American precedent in the intervention
- خلفية سوداء وسابقة أميركية في التدخل

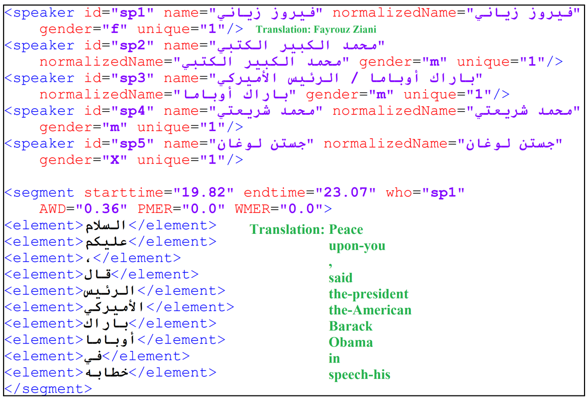
فيروز زياني: السلام عليكم، قال الرئيس الأميركى باراك أوباما فى خطابه أمام الجمعية
العامة للأمم المتحدة، انه من الضروري تجربة النهج الدبلوماسى مع إيران، تصريح
أوباما يأتى فى تناغم مع لقاء جمع بين وليام هيغ وزير خارجية بريطانيا، التى أغلقت
سفارتها في طهران منذ عامين، بنظيره الإيراني محمد جواد ظريف، كما أنها تتزامن
مع ما وصف بحملة تقارب تجاه الغرب نفذتها إيران بدءا بالإفراج عن معتقلين سياسيين
كان أوباما قد دعاها لإطلاق سراحهم وصولاً إلى تعهدات بتقديم تنازلاتٍ في الملف
النووي الإيراني.

Figure 1: *Sample input text file from Aljazeera. Output segments are underlined using different colors.*

Sample output file from QASR is shown in Figure 2. It contains speaker names as they appear in the current episode and their corresponding standardized forms across all files, which can be useful for tasks such as speaker identification and speaker linking across the entire corpus. For each speaker, we provide gender information and whether the speaker's name refers to a unique person (e.g. *Barack Obama*) or not (e.g. *One of the protesters*, or *an audio reporter*). Figure 2 has information on the anchor speaker and two guests as they appear in the metadata file, in addition to other speakers that were missed in the original transcription. It is worth noting that we provide gender and country for common Arabic speakers (who have at least 20 segments in the entire corpus). On the other hand, we ignore metadata for foreign speakers because dubbing their speeches can be done by any voice-over. We provide gender information for 2, 000 speakers and this covers 82% of all segments in the whole corpus.

Speech and text are aligned (see details in Section 2.3) and split into short segments (see Section

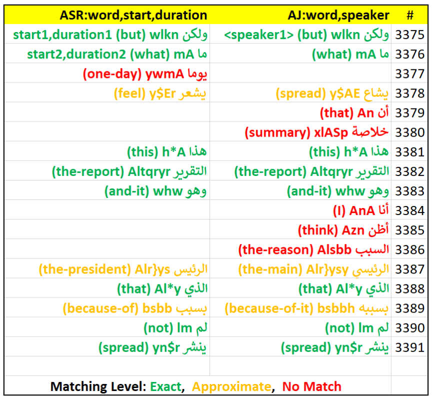
[5] We use ISO 3166 for country codes. https://en.wikipedia.org/wiki/List_of_ISO_3166_country_codes



```
<speaker id="sp1" name="فيروز زياني" normalizedName="فيروز زياني"
    gender="f" unique="1"/>  Translation: Fayrouz Ziani
<speaker id="sp2" name="محمد الكبير الكتبي"
    normalizedName="محمد الكبير الكتبي" gender="m" unique="1"/>
<speaker id="sp3" name="باراك أوباما / الرئيس الأميركي"
    normalizedName="باراك أوباما" gender="m" unique="1"/>
<speaker id="sp4" name="محمد شريعتي" normalizedName="محمد شريعتي"
    gender="m" unique="1"/>
<speaker id="sp5" name="جستن لوغان" normalizedName="جستن لوغان"
    gender="X" unique="1"/>

<segment starttime="19.82" endtime="23.07" who="sp1"
    AWD="0.36" PMER="0.0" WMER="0.0">
<element>السلام</element>      Translation: Peace
<element>عليكم</element>                    upon-you
<element>،</element>                        ,
<element>قال</element>                      said
<element>الرئيس</element>                   the-president
<element>الأميركي</element>                 the-American
<element>باراك</element>                    Barack
<element>أوباما</element>                   Obama
<element>في</element>                       in
<element>خطابه</element>                    speech-his
</segment>
```

Figure 2: *Sample output text file from QASR (XML).*



ASR:word,start,duration	AJ:word,speaker	#
start1,duration1 (but) wlkn ولكن	<speaker1> (but) wlkn ولكن	3375
start2,duration2 (what) mA ما	(what) mA ما	3376
(one-day) ywmA يوما		3377
(feel) y$Er يشعر	(spread) y$AE يشاع	3378
	(that) An أن	3379
	(summary) xlASp خلاصة	3380
(this) h*A هذا	(this) h*A هذا	3381
(the-report) Altqryr التقرير	(the-report) Altqryr التقرير	3382
(and-it) whw وهو	(and-it) whw وهو	3383
	(I) AnA أنا	3384
	(think) Azn أظن	3385
	(the-reason) Alsbb السبب	3386
(the-president) Alr}ys الرئيس	(the-main) Alr}ysy الرئيسي	3387
(that) Al*y الذي	(that) Al*y الذي	3388
(because-of) bsbb بسبب	(because-of-it) bsbbh بسببه	3389
(not) lm لم	(not) lm لم	3390
(spread) yn$r ينشر	(spread) yn$r ينشر	3391
Matching Level: Exact, Approximate, No Match		

Figure 3: *Alignment of Aljazeera transcription & ASR*

2.5). For each segment, we provide: words (*element*), timing information (*starttime* and *endtime*) in addition to speaker ID (*who*), Average Word Duration (*AWD*) in seconds, Grapheme Match Error Rate (*GWER*), and Word Match Error Rate (*WMER*). For details about word and grapheme match, refer to (Bell et al., 2015; Ali et al., 2016). Figure 2 shows information for Segment1 that appears in Figure 1.

2.3 Speech to Text Alignment

The main concept of this method is to run an Arabic speech recognition system over the entire episode (Khurana and Ali) and use the recognized word sequences and their locations in time for automatic alignment (Braunschweiler et al., 2010).

For alignment, Aljazeera and ASR transcriptions are then converted into two long sequences of words. Aligning the sequences was challenging for many reasons; code-switching between MSA and dialects; human transcription was not verbatim, e.g. some spoken words were dropped due to repetition or correction; spelling and grammar mistakes; usage of foreign languages mainly English and French; and many overlapped speeches.

We used Smith–Waterman algorithm (Smith et al., 1981), which performs local sequence alignment to determine similar regions between two strings. We modified the algorithm to accept an approximate match between the given transcription and the recognized word sequence. If the Levenshtein distance between two words $\leq$ half the length (number of characters) in the given transcription, this is considered as an approximate match.

Figure 3 shows a sample alignment, where each word is assigned to a speaker after parsing Aljazeera text and aligned, if possible, to a word from ASR transcription along with its timing information. Relaxation is applied in case of approximate match. Time information of the missing words (highlighted in red in AJ column) is estimated by interpolation from the matched word before and after. In this example, we consider words بسبب، بسببه (because-of, because-of-it) as approximate match.

2.4 Matching ASR Accuracy

Figure 4 shows the matching accuracy between the ASR and the given transcription at the segment level. We applied two levels of matching to deal with these challenges: exact match (where both transcription and ASR output are identical), and approximate match (where there is a forgiving edit distance between words in the transcription and ASR output). Exact match (100% in the *x-axis*) would have led to less than 27% of the segments, while approximate match allows to consider more segments.

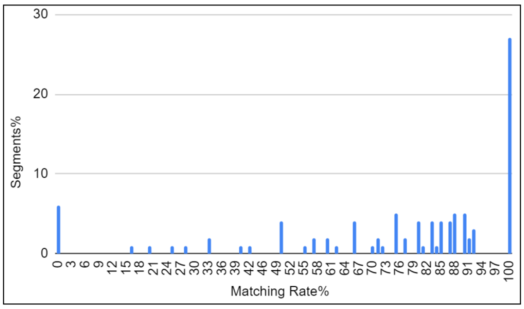


Figure 4: *Matching Accuracy between ASR and Aljazeera Transcription*

2.5 Segmentation

After aligning the given transcription with the ASR words for the whole episode, we want to segment the text into shorter segments. Unlike MGB-2, we considered many factors that we believe lead to better and logical segmentation, namely:

- **Surface:** We tried to make segments in the range of [3-10] words. We consider punctuation[6] as end of segments if they appear in this range, and we increase the window to 5 words to capture any of them in the neighbouring words. Typically, transcribers insert punctuation marks to indicate end of logical segments (sentences or phrases).
- **Dialog:** When a speaker changes in the transcribed text, we consider this as a valid end of segment. By doing this, we assign only one speaker to each segment.
- **Acoustics:** If there is a silence duration of at least 150msec between words, we consider this as a signal to potentially end the current segment. We consider the proceeding linguistic rules to confirm the validity of this end.
- **Linguistics:** For linguistically motivated segmentation, we want to avoid ending segments in wrong places (e.g. in the middle of Named Entities (NE), Noun Phrases (NP) or Adjective Phrases (AP)). To do so, from the 130M words in the LM data, we extracted the most frequent 10K words that were not followed by any punctuation in 90% of the cases, then we revised them manually[7]. We call this list ”NO-STOP-LIST”. Examples are: في، يؤدي، باتجاه (in, leads-to, towards). Additionally, we used the publicly available Arabic NLP tools (Farasa)[8] for NER and Part Of Speech (POS) tagging to label each word in the transcriptions. We put marks to avoid ending segments in the middle of NEs, NPs or APs. These are some examples from MGB-2 that have segmentation errors and words appearing erroneously in different segments: آمال /الناس (People's (seg_i) /hopes ($\text{seg}_\text{i+1}$)), مساعي /خارجية (external /endeavors) and الولايات المتحدة /الأمريكية (United States /of America). If the surface or acoustics modules suggest end of segment, while contradicting these linguistics rules, this suggestion is ignored.

Details of QASR corpus after alignment and segmentation are presented in Table 3.

2.6 Intrasentential Code-Switching

We discuss here the presence of intrasentential code-switching in QASR. We noticed in addition to the intrasentential dialectal code switching (discussed in Section 3.4), the dataset also includes $\approx 6K$ segments, where alternation between Arabic and English/French languages are seen.

CMI Range	CA	word/Utt.	#.
$0 < CMI \leq 15\%$	1.3	9.5	$1,458$
$15 < CMI \leq 30\%$	1.6	7.0	$3,806$
$30 < CMI \leq 45\%$	1.9	5.5	790
$45 < CMI \leq 100\%$	2.3	3.8	178

Table 4: Details of code-switching level in QASR data using CMI range. word/Utt. represents the average word count per utterance, CA is the mean number of code alternation points in utterances, #. presents the number of utterances for the CMI range.

To quantify the amount of code-switching present in this data, we calculate both the utterance and corpus level *Code-Mixing Index* (CMI), motivated by Chowdhury et al. (2020b); Gambäck and Das (2016). Based on the range of utterance-level CMI values, we group our dataset, as shown in Table 4. As for the corpus-level CMI, we observe an average of 30.5 CMI-value, calculated based on the average of utterance-level[9] CMI considering the code-switching segments in QASR dataset.

Furthermore, from utterance-level analysis, we notice that the majority of the code-switched segments falls under $15 < CMI \leq 30\%$ with an average of 2 alteration points per segment (e.g. Ar → En → Ar). Even though the code-switching occurs in only 0.4% of the full dataset, we notice that we have very short $\approx$ 968 segments (ranging CMI value $> 30\%$) with frequent alternating language code, such as: ”عندي duplex جوا بجنينة Building”. In the future, these segments could be used to further explore the effect of such code-switching in the performance of speech and NLP models jointly.

3 Downstream Tasks

3.1 Automatic Speech Recognition

In this section, we study QASR dataset for the ASR task. We adopt the End-to-End Transformer (E2E-T) architecture from Hussein et al. (2021) as our baseline for QASR dataset. We first augment the speech data with the speed perturbation with speed factors of 0.9, 1.0 and 1.1 (Ko et al., 2015). Then, we extract 83-dimensional feature frames consisting of 80-dimensional log Mel-spectrogram and pitch features (Ghahremani et al., 2014) and apply

[6] Common punctuation marks are: Period, Comma, Question mark, Exclamation mark, Semicolon, Colon and Ellipsis.

[7] The final list has 2,200 words.

[8] farasa.qcri.org

[9] Excluding switches between the utterances.

	Dev_WER /[S, D, I]	Test_WER /[S, D, I]
Best E2E-T-MGB-2	15.0 [10.0, 3.9, 1.1]	14.3 [9.5, 3.7, 1.1]
Baseline E2E-T-QASR	15.1 [7.0, 7.4, 0.7]	14.7 [7.1, 7.0, 0.6]

Table 5: *WER% performance with the insertion (I%), substitution (S%) and deletion (D%) rates for the transformer ASR (E2E-T) pretrained on QASR and MGB-2.*

cepstral mean and variance normalization. Furthermore, we augment these features using the specaugment approach (Park et al., 2019). We use Espnet (Watanabe et al., 2018) to train the E2E-T model on MGB-2 and QASR datasets. Each model was trained for 30 epochs using 4 NVIDIA Tesla V100 GPUs, each with 16 GB memory, which lasted two weeks. Results of the baseline model on both development and testsets are shown in Table 5.

It can be seen that the best E2E-T-MGB-2 achieves slightly better WER with a difference of 0.3% on average. This is expected since adopted E2E-T architecture was carefully tuned on MGB-2 dataset. However, the E2E-T-QASR achieves lower substitution and insertion rates with an absolute difference of 2.7% and 0.5% on average respectively. It can also be noticed that almost half of the E2E-T-QASR errors are due to deletions. To investigate these results further, we visualize the distribution of segmentation duration of the MGB-2 train, the QASR train and the testsets as shown in Figure 5. We consider the range within 3 standard deviations of each distribution as the effective segmentation duration that contains 99% of the segments, and the rest 1% of the segments are considered as outliers. From Figure 5, it can be seen that QASR distribution is following the bell curve similar to the testset which was segmented by an expert transcriber. On the other hand, the MGB-2 distribution is right-skewed with segment duration outliers that go beyond 50 seconds. In addition, one can observe that the effective segmentation duration of the testset is 9 seconds, which is larger than QASR effective segmentation duration, which is only 7 seconds. On the other hand, the MGB-2 effective segmentation duration covers a much larger range of over 30 seconds. The difference in the segment duration affects the statistical properties of the data and causes a shift in the data distribution. We think that this is the main reason why the baseline E2E-T-QASR achieves worse results than best E2E-T-MGB-2. To validate our assumption, we analyze the E2E-T-QASR transcription and found that the deletion errors mainly appeared with segments that are larger than 7 seconds. We illustrate our findings with two transcription examples in Buckwalter (BW) format shown in Figure 6: short segment of 6 seconds, and long segment of 10 seconds. Deletions are highlighted in red, substitutions in yellow, and correct in green. It can be seen from the short example that E2E-T-QASR achieves better results with a potential for code-switching. On the other hand, the long example confirms our assumption about the shift in segments duration distribution between QASR and the testset.

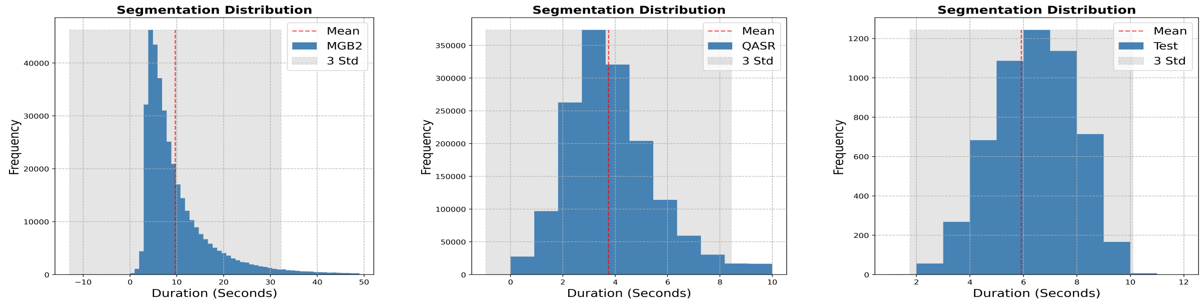
3.2 Automatic Punctuation Restoration

Cl.	QASR	Dev	Test	Fisher
،	428K (3.2%)	2K (3.0%)	1K (2.5%)	70K (11.8%)
.	154K (1.2%)	1K (2.5%)	1K (1.8%)	362K (6.3%)
؟	87K (0.7%)	623 (0.9%)	349 (0.5%)	56K (1.3%)
O	12M (95.0%)	68K (93.6%)	63K (95.1%)	2M (80.6%)

Table 6: *Distribution of punctuation classes in QASR (Arabic) along with 348 hours of Fisher (English) corpus as a reference. O representing – No Punctuation, COMMA (،), FSTOP (.), Ques (؟).*

In this section, we explore QASR for the automatic punctuation restoration task. To prepare the training data, we first segment the utterances from the same speaker with a maximum window of 120 tokens. We then remove utterances with ≤ 6 words and no punctuation in the segment. We pre-process the lexical utterances, removing diacritics, brackets, among others. For the task, we only keep the top 3 punctuation classes ('،', '؟' and '.') and rest are mapped to class 'O' representing *no punctuation*. The distribution of punctuation in QASR are highly imbalanced (as shown in Table 6), which is expected of a spoken corpus. However, in comparison to the Fisher corpus (Cieri et al., 2004) and other language datasets (see (Li and Lin, 2020)), the distribution is more skewed. This is because in Arabic, punctuation marks are rarely used, e.g., *Segment1* in Figure 1, can be logically divided into two segments separated by a full stop.

We adapt a simple transformer-biLSTM architecture (Alam et al., 2020) as our baseline model using lexical information. Given an input token sequence $(x_1, x_2..., x_m)$, we extract the subwords $(s_1, s_2..., s_n)$ using wordpiece tokenizer. These subwords are fed into the pre-trained BERT model, which outputs a vector of d dimension for each time step. These d vectors are then passed to a BiLSTM

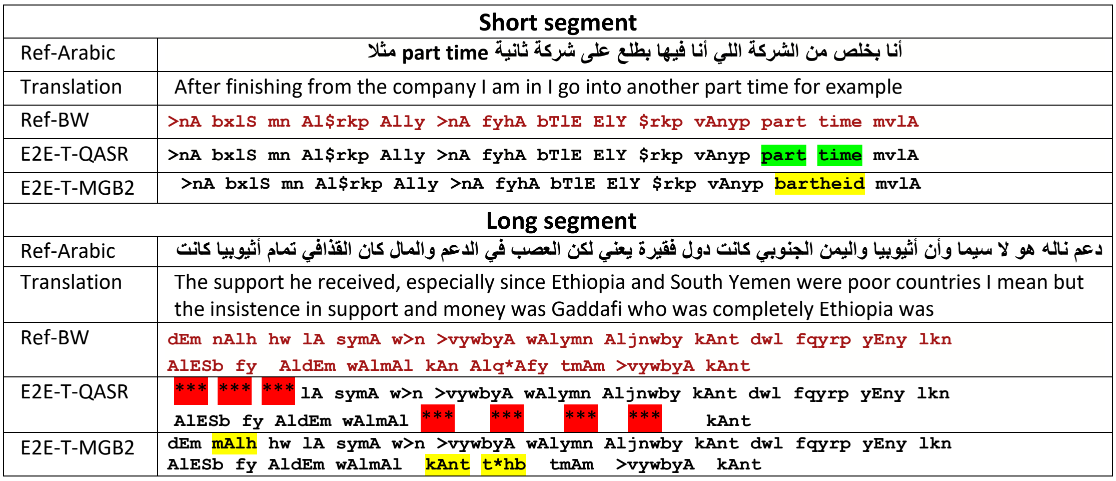


(a)

(b)

(c)

Figure 5: *Duration distributions of the speech segments for: (a) MGB-2, (b) QASR, and (c) test dataset.*



Short segment	
Ref-Arabic	أنا بخلص من الشركة اللي أنا فيها بطلع على شركة ثانية part time مثلا
Translation	After finishing from the company I am in I go into another part time for example
Ref-BW	>nA bxlS mn Al$rkp Ally >nA fyhA bTlE ElY $rkp vAnyp part time mvlA
E2E-T-QASR	>nA bxlS mn Al$rkp Ally >nA fyhA bTlE ElY $rkp vAnyp part time mvlA
E2E-T-MGB2	>nA bxlS mn Al$rkp Ally >nA fyhA bTlE ElY $rkp vAnyp bartheid mvlA
Long segment	
Ref-Arabic	دعم ناله هو لا سيما وأن أثيوبيا واليمن الجنوبي كانت دول فقيرة يعني لكن العصب في الدعم والمال كان القذافي تمام أثيوبيا كانت
Translation	The support he received, especially since Ethiopia and South Yemen were poor countries I mean but the insistence in support and money was Gaddafi who was completely Ethiopia was
Ref-BW	dEm nAlh hw lA symA w>n >vywbyA wAlymn Aljnwby kAnt dwl fqyrp yEny lkn AlESb fy AldEm wAlmAl kAn Alq*Afy tmAm >vywbyA kAnt
E2E-T-QASR	*** *** *** lA symA w>n >vywbyA wAlymn Aljnwby kAnt dwl fqyrp yEny lkn AlESb fy AldEm wAlmAl *** *** *** *** kAnt
E2E-T-MGB2	dEm mAlh hw lA symA w>n >vywbyA wAlymn Aljnwby kAnt dwl fqyrp yEny lkn AlESb fy AldEm wAlmAl kAnt t*hb tmAm >vywbyA kAnt

Figure 6: E2E-T-QASR and E2E-T-MGB-2 transcription on short segment and long segments of 6 and 10 seconds respectively. Each example includes text in Arabic, Buckwalter (BW) and English translation.

Dev	O	COMMA	FSTOP	QUES
P	97.0%	52.7%	78.8%	61.3%
R	98.8%	38.8%	50.4%	59.7%
F1	97.9%	44.7%	61.5%	60.5%
Test	O	COMMA	FSTOP	QUES
P	98.1%	44.9%	70.7%	52.6%
R	98.4%	48.6%	51.7%	57.3%
F1	98.3%	46.7%	59.7%	54.9%

Table 7: *Reported Precision (P), Recall (R) and F-measure (F1) on test and dev set using punctuation restoration model trained on QASR dataset.*

layer, consisting of h hidden units. The choice of using BiLSTM is to make effective use of both past ($\overrightarrow{h}$) and future ($\overleftarrow{h}$) contexts for prediction. The concatenated $\overrightarrow{h} + \overleftarrow{h}$ output at each time step is then fed to a fully-connected layer with four output neurons, which correspond to 3 punctuation marks and the 'O' token.

During the training, special tokens identifying start- and end-of the sentence are added to the input subword sequence.[10] For this task, we used AraBERT (Antoun et al., 2020): pre-trained on newspaper articles, containing 3 transformer self attention layers with each hidden layer of 768. These token embeddings are then passed onto a BiLSTM with hidden dimension of 768. The baseline model is trained using Adam optimizer with a learning rate of $1e-5$ and 32 batch size for 10 epochs.

Despite the fact that Arabic has a skewed distribution in punctuation, the baseline results reported in Table 7 for the 3 punctuation and 'O' labels show that the prediction results of the full stop and the question mark are better than the comma. This again reconfirms that in Arabic, the use of comma is highly debatable (Mubarak et al., 2015; Mubarak and Darwish, 2014) and can easily be substituted by the full stop or other punctuation. In the future, we will explore better architectures with information from different modalities, such as acoustics.

[10] The maximum length of the subwords is set to 256. In cases, if the sequence exceeds the maximum length, it is then divided into two separate sequences.

Speakers – Male – Female	40 (Anchor (A): 20, Guest (G): 20) 33 (A: 14, G:19) 7 (A:6, G:1)
Segments	$4,000$ (100 / speaker)
Countries	11 unique countries (DZ, EG, IQ, LB, LY, MA, PS, SA, SY, TN, YE)

Table 8: *QASR subset used for speaker verification (SV) and Arabic dialect identification (ADI) tasks.*

3.3 Speaker Verification

One of the biggest challenges in broadcast domain is its speech diversity. The anchor speaker voice is often clear and planned. However, the spoken style[11] of different program guests can present various challenges. Here, we showcase how QASR could be used to evaluate existing speaker models based on the speakers' role in each episode. In the future, the dataset can also be used to study turn-taking and speaker dynamics, given the interaction between speakers in QASR.

Sets	EER	Total Pairs
Anchor	9.2	$40K$ (75% male)
Guest	7.5	$40K$ (100% male)
Mixed	7.9	$40K$ (75% male)
VoxCeleb1-tst	6.8	38K

Table 9: *Reported EER on verification trial pairs for anchors, guest and their combination. In addition, EER reported on VoxCeleb1 official test verification pairs (English) as reference. 50% of the total pairs are positive, i.e. from same speaker.*

We adapt one of the widely-known architectures used to model an end-to-end text-independent Speaker Recognition (SR) system. For the study, we use a pre-trained model, with four temporal convolution neural networks followed by a global (statistical) pooling layer and then two fully connected layers. The input to the model is MFCCs features (with 40 coefficient) computed with a 25msec window and 10ms frame-rate from the 16KHz audio. The model is trained on Voxceleb1 (Nagrani et al., 2017) development set (containing $1,211$ speakers and $\approx$ 147K utterances). More details can be found in Shon et al. (2018); Chowdhury et al. (2020a).

For speaker verification, we use verified same/different-speaker pairs of speech segments as input. We extract the length normalized embeddings from the last layer of the SR model and then computed the cosine similarity between pairs.

For our evaluation, we constructed these verification pair trials by randomly picking up $40K$ utterance pairs from: (*i*) speakers of the same gender; (*ii*) similar utterance lengths; and (*iii*) a balanced distribution between positive and negative targets[12]. For this, we use the most frequent 20 anchor and 20 guest speakers data subset described in Table 8. We then compare the Equal Error Rate (EER) of the model, reported in Table 9, using the designed verification pairs based on a particular job role, or their combination. In addition, we also report the results on VoxCeleb1 official verification testset as a reference.

From the results, we observe that the SR model effectively distinguishes between the positive and negative pairs with $\approx$ 70% (A) - 72% (G) accuracy. Comparing the EER, we notice that it is harder to differentiate between anchors than guests. This can be due to the fact that anchors are using the same acoustic conditions, and the current models are learning recording conditions (Chowdhury et al., 2020a) as well as speaker information.

3.4 Arabic Dialect Identification

To understand the dialectal nature of QASR dataset, we analyze the acoustic and lexical representations for 100 segments from each speaker[13].

To obtain the dialect labels, we run the pre-trained dialect identification models for both speech and text modality. We address the dialect identification as multi-stage classification: Firstly, we predict the labels of the segments - MSA *vs* DA - and, secondly, if the label is DA, we further propagate the labels to detect the country of the selected speaker (i.e fine-grained dialect classification). For country level evaluation, we manually annotate each speaker's country label (see Table 8).

For lexical modality, we use the pre-trained QADI (Abdelali et al., 2020), and for the acoustic modality, we use ADI-5[14] (Shon et al., 2017; Ali et al., 2019) – as MSA vs DA classifier – along with ADI-17[15] (Shon et al., 2020) for fine-grained labels.

[11] The style can vary based on language fluency, speech rate, use of different dialects among other factors.

[12] The official verification pairs are included as a part of QASR.

[13] We used the same speaker set as the SV task.

[14] https://github.com/swshon/dialectID_e2e

[15] https://github.com/swshon/arabic-dialect-identification

We observe that in both the modalities, 50% of the anchors speak MSA in 70% of the time in speech and 90% of the time in text. As for the other 50%, we notice that using the dialect identification modules, we can detect only 20% of the speaker's nationality correctly. The aforementioned observations are pre-anticipated, as anchors are professionally trained to speak mostly in MSA, making it harder for the model to predict the correct country label. This also explains why the large portion of the data is MSA.

As for guest speakers, we notice that the lexical classifier detected that 30% of the speakers use MSA, while 70% of the speakers were detected as DA. As for the acoustic models, we notice that all speakers use dialects more than 70% of the time. Comparing the accuracy of identifying the correct dialects based on annotated country labels, we notice that both the text and acoustic models perform comparatively better in identify the guest speakers' country – 64% from text and 65% from acoustic. Our hypothesis for such increase in performance is that guest speakers, unlike the anchors, mostly speak using their dialects, making it easier for the model to infer their country.

When comparing the decision from both modalities, we notice that there is an agreement of 67.5% (65% for anchor and 70% for guest speakers) for MSA/DA classification. Most of the classification errors in speech and text dialect identification models are due to confusion between dialects spoken in neighboring countries; e.g. Syria and Lebanon in the Levantine region; Tunisia and Algeria in the North African region.

3.5 Named Entity Recognition (NER)

NER is essential for a variety of NLP applications such as information extraction and summarization. There are many researches on Arabic NER for news articles, e.g. ANERcorp (Benajiba and Rosso, 2008) and microblogs (Darwish, 2013). However, we are not aware of any studies or datasets for NER in Arabic news transcription, which can be useful for applications like video search. We manually annotate and revised the MGB-2 testset for basic NE types, namely Person (PER), Location (LOC), Organization (ORG) and Others (OTH/MISC) following the guidelines in (Benajiba and Rosso, 2008). The testset (70K words) along with NER annotation is available as part of QASR. From the annotation, we observed NEs are 7% of the corpus and their distribution is as follows: PER= 32%, LOC= 46%, ORG= 18% and OTH= 5%[16].

	ANERcorp			QASR		
Type	P	R	F1	P	R	F1
PER	87.0	77.7	82.1	62.8	51.2	56.4
LOC	92.3	87.8	90.0	86.4	88.1	87.2
ORG	81.4	66.0	72.9	22.8	19.1	20.8
Overall	88.7	80.3	84.3	72.2	67.5	69.8

Table 10: NER results: Precision (P), Recall (R) and F1 on two testsets.

We test the publicly available Arabic Farasa NER on our new testset and compare performance with the standard news testset (ANERcorp). Results are listed in Table 10. As shown, testing NER on transcribed speech has lower F1 by 15% compared to testing on a standard news testset (from 84.3% to 69.8%). We anticipate that characteristics of speech transcription described in Section 2.3 affected NER negatively[17]. We keep enhancing NER for speech transcription for future work.

4 Conclusion

In this paper, we introduce a $2,000$ hours transcribed Arabic speech corpus, QASR. We report results for automatic speech recognition, Arabic dialect identification, speaker verification, and punctuation restoration to showcase the importance and usability of the dataset. QASR is also the first Arabic speech-NLP corpus to study spoken modern standard Arabic and dialectal Arabic. We report for the first time named entity recognition in Arabic news transcription. The $11,092$ unique speakers present in QASR can be used to study turn-taking and speaker dynamics in the broadcast domain. The corpus can also be useful for unsupervised methods to select speaker for text to speech (Gallegos et al., 2020). The QASR is publicly available for the research community.

Acknowledgements

This work was made possible with the collaboration between Qatar Computing Research Institute, HBKU and Aljazeera media network. This data is hosted on ArabicSpeech portal[18], which is a community based effort that runs for the benefit of Arabic speech science and technologies.

[16] ANERcorp contains 150K words. NEs are 11%. Distribution: PER= 39%, LOC= 30%, ORG= 21%, MISC= 10%.

[17] Disfluency example: طيب أنتم أليست يعني ألا تستحون؟ (OK you are isn't it I mean are you not ashamed?)

[18] https://arabicspeech.org/

Ethical Concern and Social Impact

User Privacy

QASR dataset only includes programs that have been broadcast by the Aljazeera news media. No additional identity of the guest is revealed in the data, which was made anonymous in the original program. However, in the future, if any concern is raised for a particular content, we will comply to legitimate concerns by removing the affected content from the corpus.

Biases in QASR

Any biases found in the dataset are unintentional, and we do not intend to do harm to any group or individual. The bias in our data, for example towards a particular gender is unintentional and is a true representation of the programs. We do address these concerns by collecting examples from both parties before any general suggestion.

As for the assigned annotation label, we follow a well-defined schema and available information to perceive a final label. For e.g. gender label – male/female is perceived from the data and might not be a true representative of the speakers' choice.

Potential Misuse

We request the research community to be aware that our dataset can be used to misuse quotes for the speakers for political or other gain. If such misuse is noticed, human moderation is encouraged in order to ensure this does not occur.

References

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa: A fast and furious segmenter for Arabic. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Demonstrations*, pages 11–16.

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2020. Arabic dialect identification in the wild. *arXiv preprint arXiv:2005.06557*.

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. Nadi 2020: The first nuanced Arabic dialect identification shared task. *arXiv preprint arXiv:2010.11334*.

Tanvirul Alam, Akib Khan, and Firoj Alam. 2020. Punctuation restoration using transformer models for resource-rich and-poor languages. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 132–142.

Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang. 2016. The MGB-2 challenge: Arabic multi-dialect broadcast media recognition. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 279–284. IEEE.

Ahmed Ali, Shammur Chowdhury, Mohamed Afify, Wassim El-Hajj, Hazem Hajj, Mourad Abbas, Amir Hussein, Nada Ghneim, Mohammad Abushariah, and Assal Alqudah. 2021. Connecting Arabs: bridging the gap in dialectal speech recognition. *Communications of the ACM*, 64(4):124–129.

Ahmed Ali, Suwon Shon, Younes Samih, Hamdy Mubarak, Ahmed Abdelali, James Glass, Steve Renals, and Khalid Choukri. 2019. The MGB-5 challenge: Recognition and dialect identification of dialectal Arabic speech. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1026–1033. IEEE.

Ahmed Ali, Stephan Vogel, and Steve Renals. 2017. Speech recognition challenge in the wild: Arabic MGB-3. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 316–322. IEEE.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for Arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Peter Bell, Mark JF Gales, Thomas Hain, Jonathan Kilgour, Pierre Lanchantin, Xunying Liu, Andrew McParland, Steve Renals, Oscar Saz, Mirjam Wester, et al. 2015. The MGB challenge: Evaluating multi-genre broadcast media recognition. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 687–693. IEEE.

Yassine Benajiba and Paolo Rosso. 2008. Arabic named entity recognition using conditional random fields. In *Proc. of Workshop on HLT & NLP within the Arabic World, LREC*, volume 8, pages 143–153. Citeseer.

Norbert Braunschweiler, Mark JF Gales, and Sabine Buchholz. 2010. Lightly supervised recognition for automatic alignment of large coherent speech recordings. In *Eleventh Annual Conference of the International Speech Communication Association*.

Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina, et al. 2018. State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4774–4778. IEEE.

Shammur A Chowdhury, Ahmed Ali, Suwon Shon, and James Glass. 2020a. What does an end-to-end dialect identification model learn about non-dialectal information? *Proc. Interspeech 2020*, pages 462–466.

Shammur A Chowdhury, Younes Samih, Mohamed Eldesouki, and Ahmed Ali. 2020b. Effects of dialectal code-switching on speech modules: A study using Egyptian Arabic broadcast speech. *Proc. Interspeech*.

Shammur Absar Chowdhury, Amir Hussein, Ahmed Abdelali, and Ahmed Ali. 2021. Towards one model to rule all: Multilingual strategy for dialectal code-switching Arabic Asr. *arXiv:2105.14779*.

Shammur Absar Chowdhury, Evgeny Stepanov, Morena Danieli, and Giuseppe Riccardi. 2017. Functions of silences towards information flow in spoken conversation. In *Proceedings of the Workshop on Speech-Centric Natural Language Processing*, pages 1–9.

Shammur Absar Chowdhury, Evgeny A Stepanov, Morena Danieli, and Giuseppe Riccardi. 2019. Automatic classification of speech overlaps: Feature representation and algorithms. *Computer Speech & Language*, 55:145–167.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The fisher corpus: a resource for the next generations of speech-to-text. In *LREC*, volume 4, pages 69–71.

Kareem Darwish. 2013. Named entity recognition using cross-lingual resources: Arabic as an example. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1558–1567.

Chenpeng Du, Hao Li, Yizhou Lu, Lan Wang, and Yanmin Qian. 2021. Data augmentation for end-to-end code-switching speech recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 194–200. IEEE.

Pilar Oplustil Gallegos, Jennifer Williams, Joanna Rownicka, and Simon King. 2020. An unsupervised method to select a speaker subset from large multi-speaker speech synthesis datasets. *Proc. Interspeech 2020*, pages 1758–1762.

Björn Gambäck and Amitava Das. 2016. Comparing the Level of Code-Switching in Corpora. In *LREC*.

Pegah Ghahremani, Bagher BabaAli, Daniel Povey, Korbinian Riedhammer, Jan Trmal, and Sanjeev Khudanpur. 2014. A pitch extraction algorithm tuned for automatic speech recognition. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2494–2498. IEEE.

Sabit Hassan, Hamdy Mubarak, Ahmed Abdelali, and Kareem Darwish. 2021. Asad: Arabic social media analytics and understanding. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 113–118.

Amir Hussein, Shinji Watanabe, and Ahmed Ali. 2021. Arabic speech recognition by end-to-end, modular systems and human. *arXiv preprint arXiv:2101.08454*.

Sameer Khurana and Ahmed Ali. QCRI advanced transcription system (QATS) for the Arabic multi-dialect broadcast media recognition: MGB-2 challenge. *Training*, 1200(2214):370K.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Xinxing Li and Edward Lin. 2020. A 43 language multilingual punctuation prediction neural network model. *Proc. Interspeech 2020*, pages 1067–1071.

Hamdy Mubarak, Ahmed Abdelali, Hassan Sajjad, Younes Samih, and Kareem Darwish. 2019. Highly effective Arabic diacritization using sequence to sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2390–2395.

Hamdy Mubarak and Kareem Darwish. 2014. Automatic correction of Arabic text: A cascaded approach. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 132–136.

Hamdy Mubarak, Kareem Darwish, and Ahmed Abdelali. 2015. Qcri@ qalb-2015 shared task: Correction of Arabic text for native and non-native speakers' errors. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 150–154.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*.

Joseph Olive, Caitlin Christianson, and John McCary. 2011. *Handbook of natural language processing and machine translation: DARPA global autonomous language exploitation*. Springer Science & Business Media.

David S Pallett. 2003. A look at nist's benchmark asr tests: past, present, and future. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)*, pages 483–488. IEEE.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *Proc. Interspeech*, pages 2613–2617.

Arfath Pasha, Mohamed Al-Badrashiny, Mona T Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth.

2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *LREC*, volume 14, pages 1094–1101. Citeseer.

Sheena Christabel Pravin and M Palanivelan. 2021. A hybrid deep ensemble for speech disfluency classification. *Circuits, Systems, and Signal Processing*, pages 1–28.

Younes Samih, Mohammed Attia, Mohamed Eldesouki, Ahmed Abdelali, Hamdy Mubarak, Laura Kallmeyer, and Kareem Darwish. 2017. A neural architecture for dialectal Arabic segmentation. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 46–54.

George Saon, Gakuto Kurata, Tom Sercu, Kartik Audhkhasi, Samuel Thomas, Dimitrios Dimitriadis, Xiaodong Cui, Bhuvana Ramabhadran, Michael Picheny, Lynn-Li Lim, et al. 2017. English conversational telephone speech recognition by humans and machines. *arXiv preprint arXiv:1703.02136*.

Suwon Shon, Ahmed Ali, and James Glass. 2017. MIT-QCRI Arabic dialect identification system for the 2017 multi-genre broadcast challenge. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 374–380. IEEE.

Suwon Shon, Ahmed Ali, Younes Samih, Hamdy Mubarak, and James Glass. 2020. Adi17: A fine-grained Arabic dialect identification dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8244–8248. IEEE.

Suwon Shon, Najim Dehak, Douglas Reynolds, and James Glass. 2019. Mce 2018: The 1st multi-target speaker detection and identification challenge evaluation. *arXiv preprint arXiv:1904.04240*.

Suwon Shon, Hao Tang, and James Glass. 2018. Frame-level speaker embeddings for text-independent speaker recognition and analysis of end-to-end model. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 1007–1013. IEEE.

Temple F Smith, Michael S Waterman, et al. 1981. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197.

Anshuman Tripathi, Han Lu, and Hasim Sak. 2020. End-to-end multi-talker overlapping speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6129–6133. IEEE.

Jörgen Valk and Tanel Alumäe. 2020. Voxlingua107: a dataset for spoken language recognition. *arXiv preprint arXiv:2011.12998*.

Changhan Wang, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv:2101.00390*.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson-Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al. 2018. Espnet: End-to-end speech processing toolkit. *Proc. Interspeech*, pages 2207–2211.

Jane Wottawa, Marie Tahon, Apolline Marin, and Nicolas Audibert. 2020. Towards interactive annotation for hesitation in conversational speech. In *LREC 2020*.

Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. 2016. Achieving human parity in conversational speech recognition. *arXiv preprint arXiv:1610.05256*.